Data Hiding Domain Classification for Blind Image Steganalysis

Guo-Shiang Lin¹, Chia H. Yeh² and C.-C. Jay Kuo²

¹Department of Electrical Engineering, National Chung Cheng University, Chia-Yi, 621 Taiwan
²Integrated Media Systems Center and Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089-2564
E-mail: khlin@samlab.ee.ccu.edu.tw, {chyeh,cckuo}@sipi.usc.edu

Abstract

A statistical feature-based scheme is proposed to identify the data hiding domain of an embedded signal in this research. Two phenomena are observed for images before and after data hiding. First, the gradient energy increases as the continuity of gray levels between adjacent pixels is disturbed by the embedded signal. Second, the statistical variance of the coefficient distribution in macro-blocks tends to decrease after data hiding. These phenomena are analyzed mathematically. Then, statistical features in the pixel, DCT, and DWT domains are extracted and a maximum likelihood ratio test is adopted to solve the hiding domain classification problem. The proposed scheme has demonstrated good classification results.

1. Introduction

Data hiding has been an active research field in recent years. One area of interest is known as “steganalysis”, which can be used to enhance the capability of warders for detecting hidden data (i.e. passive steganalysis) or altering, extracting, and tracking the hidden data to thwart any secret communication (i.e. active steganalysis) [2]. For image steganalysis, one may detect suspicious objects and extract hidden messages by comparing them with original versions. However, due to the constraints of portability and limited accessibility to original versions, blind image steganalysis is often demanded in practical applications.

Steganalytic algorithms were developed to analyze spatial-domain steganographic schemes. Fridrich et al. [1] proposed a steganalysis technique by exploring the fact that bit planes in typical images are more or less correlated so that the LSB plane can be estimated from the other 7 higher level bit planes. This estimation becomes less reliable if the content of the LSB bit plane is randomized before casting. Avcibas et al. [2] claimed that any image would incur quality degradation after smoothing or low-pass filtering operations, and this degradation (reacting on image quality) depends on the type of the test image, especially in categories of with or without embedded information. That is, by observing quality difference between a test image and its smoothed version, it is possible to discriminate images with and without hidden messages.

Steganalytic methods for data hiding in the DCT domain were also studied. Fridrich et al. [1] claimed that a modified image block would most likely become saturated (i.e. at least one pixel with the grey value 0 or 255) in a JPEG-format stego-image after data hiding. Thus, if no saturated blocks can be found, one may claim that no secret messages exist. The spatial-domain steganalytic method [1] mentioned earlier can also be used to analyze these saturated blocks.

Generally speaking, hidden data can be cast in the spatial, the DCT, and the DWT (discrete wavelet transform) domains, and steganographic schemes with different robustness capabilities may be exploited to resist the passive or active steganalysis in different applications. Therefore, after hidden data detection for a given stego image, we are interested in finding out the original data hiding domain because the hiding domain is a useful “clue” for the active steganalysis. Statistical features are extracted in the pixel, DCT, and DWT domains and a maximum likelihood ratio test is performed to achieve this goal in this research.

2. Hiding Domain Classification

Data embedded in an image will alter insignificant parts of data and introduce distortions, which could be hardly observed by human eyes but tend to change properties of the host image. Some statistical properties for hidden message detection were described before, e.g. randomization of the LSB plane, gray-level changes between groups of pixels, and so on. We consider features in the pixel, DCT, and DWT domains for hiding domain classification as follows.

2.1 Feature Selection in Three Domains

Given a steganographic image, we examine it from the following domains: pixel, DCT, and DWT domains. The effect of data hiding in the pixel and the DCT domains is localized. Thus, the block-based local processing is adopted in the pixel and DCT domains. Our strategy is to select macro-blocks (MBs) with a low gradient value as candidates for steganalysis. Here, the size of MB is 16×16 pixels, the Sobel operator is
adopted to calculate the gradient, and one-third of all MBs are chosen as candidate MBs.

**Pixel domain.** After candidate MBs are chosen, we calculate the average of gradients of all candidate MBs as a feature denoted by $f_r^p$. Besides, we only calculate the energy of the middle and high bands in each block because the secret data is often encrypted to be a noisy-like signal before information hiding. The average of energies is used as the second feature $f_z^p$. Then, we have a 2D feature vector $f^p = (f_r^p, f_z^p)$.

**DCT domain.** In the DCT domain, we model AC coefficients in an MB to be Laplacian-distributed. The parameter that is estimated via maximum likelihood [3] forms a feature, which is denoted by $f_i^d$. Furthermore, we compute the average of energies of middle and high bands in selected MBs as the second feature $f_z^d$. Thus, we obtain a 2D feature vector $f^d = (f_i^d, f_z^d)$.

**DWT domain.** In the DWT domain, we divide high frequency subbands into several non-overlapped MBs and obtain all coefficient histograms. After calculating the standard deviation (STD) of the histogram in each MB, we can compute the average of STDs as a feature $f_i^w$. Besides, the average of gradients along both vertical and horizontal directions is calculated as the second feature $f_z^w$. Then, we are led a 2D feature vector $f^w = (f_i^w, f_z^w)$.

### 2.2 Proposed Classification Scheme

We propose a two-stage classification scheme to classify the data hiding domain. In the first stage, after all features are computed, whether a test image conveys the hidden information or not in a given domain is a two-class classification problem. The subsequent issue is to find a good classifier that maximizes the classification rate. This is a straightforward problem which has been widely discussed.

It is observed that data embedding in the pixel and the DCT domains have similar characteristics so that stego-images with data hidden in the DCT and the pixel domains are difficult to distinguish. Thus, we focus on the problem arising in the second stage. We first examine whether DWT is the data hiding domain. If we see a positive response in the DWT domain, it is claimed to be a DWT domain embedding. If not, then we develop a scheme to perform hiding domain discrimination between DCT and pixel domains, which is the main contribution of our current work.

The basic idea is to lessen the embedded signal in the pixel domain and then measure the change of the extracted feature for classification. To reduce the embedded signal and obtain the reference, the estimated version can be achieved based on the watermark removal technique [4]. After the averaged gradient along vertical and horizontal directions is calculated, a ratio of the gradient of the test image to that of its corresponding estimated reference can be computed as a feature. A multi-layer neural network is then employed for classification.

### 3. Statistical Feature Analysis

In Section 2, we introduce features to be selected in three domains. Basically, they are related to either the gradient energy or the coefficient distribution. In this section, we analyze how these features change due to information hiding. Before our discussion, we first give some notations.

1. $I^r$ and $t^r$ represent the original image and the stego version, respectively.
2. $r^p$ and $t^p$ stand for the original coefficient and the stego version, respectively.
3. $A$ is a set of pixels or coefficients selected by a steganographic key and $\overline{A}$ is another set except $A$.
4. $N_i$ and $N_2$ stand for the sizes of $A$ and $\overline{A}$, respectively.
5. $w$ represents the transmitted information (a binary, +1/-1, or a real number) and $\alpha$ is a scale factor.

#### 3.1 Gradient energy modification

Based on energy preservation, regardless of which domain is adopted to embed secret data, a general rule in the spatial domain can be expressed as

$$I^r = I^t + \alpha \cdot w \text{ OF } I^r = I^t + (1 + \alpha \cdot w).$$

(1)

Discussion on the gradient energy change is first given for the one-dimensional (1-D) case. Considering a 1-D sequence $I(n)$, gradient $\gamma_r(n)$ between two successive points within $A$ before data hiding can be written as

$$\gamma_r(n) = I'_r(n) - I'_r(n-1).$$

(2)

The averaged gradient energy in $A$ becomes

$$GE^A = \frac{1}{N_i - 1} \sum \gamma_r^2 = \frac{1}{N_i - 1} \sum (I'_r(n) - I'_r(n-1))^2.$$  

(3)

Considering $GE^A$ as a random variable (a function of $I'_r$), the expectation of the gradient energy can be calculated as:

$$E[GE^A] = \frac{1}{N_i - 1} \sum E[\gamma_r^2].$$  

(4)

where $E[\cdot]$ stands for the expectation of a random variable. After information hiding, $I'(n)$ changes to $I'(n)$ and the gradient function can be written as
\[ r^h_\alpha(n) = I^h_\alpha(n) - I^{h-1}_\alpha(n-1) = q(n) - q(n-1) \] (5)

where \( q(n) \) represents the disturbing sequence (i.e., \( q = \alpha \cdot w \) or \( I^h \cdot \alpha \cdot w \) in Eq. (1)) hidden in \( I^\alpha(n) \) and \( r^h_\alpha(n) \) is similarly defined as \( r^h_\alpha(n) = q(n) - q(n-1) \). Then, the expectation of the averaged gradient energy of \( I^\alpha(n) \) in \( A \) is

\[ E[GE^\alpha_n] = E\left[\frac{1}{N_1-1} \sum_{w} (r^\alpha_w(n))^2\right] \]

\[ = \frac{1}{N_1-1} \sum_{w} E[(r^\alpha_w(n))^2] + 2 \cdot r^\alpha_w(n) - r^\alpha_w(n) + (r^\alpha_\alpha(n))^2]. \] (6)

In most data hiding schemes, secret messages are often scrambled before the embedding procedure. In addition, the creation of secret messages is generally independent of the cover-image. Thus, it is reasonable to assume that \( r^\alpha_w \) is statistically independent of \( r^\alpha_\alpha \), i.e.,

\[ E[r^\alpha_w(n) - r^\alpha_\alpha(n)] = E[r^\alpha_w(n)] - E[r^\alpha_\alpha(n)]. \]

Since \( r^\alpha_w \) can be considered as a prediction error of \( I^\alpha(n) \) from \( I^{\alpha-1}(n-1) \) within \( A \), which was often modeled as a Gaussian or Laplacian distribution with zero mean, i.e., \( E[r^\alpha_w(n)] = 0 \), we can obtain \( E[(r^\alpha_w(n))^2] = 0 \) and then

\[ E[GE^\alpha_n] = \frac{1}{N_1-1} \sum_{w} E[(r^\alpha_w(n))^2] = E[GE^\alpha_n]. \] (7)

In the same manner, because pixels are innocent in \( \bar{A} \), we can obtain \( GE^\bar{A}_\alpha = GE^\bar{A}_\alpha \), where \( GE^\alpha \) and \( GE^\bar{A} \) are gradient energies of the original set and the stego one within \( \bar{A} \), respectively. Then, we can obtain

\[ E[GE^\alpha] = \frac{N_1}{N_1+N_2} E[GE^\alpha] + \frac{N_2}{N_1+N_2} E[GE^\bar{A}] \]

\[ = E[GE^\alpha] + \frac{N_1}{N_1+N_2} \cdot \frac{1}{N_1-1} \sum_{w} E[(r^\alpha_w(n))^2]. \] (8)

Since \( E[(r^\alpha_w(n))^2] \geq 0 \), we obtain

\[ E[GE^\alpha] \geq E[GE^\alpha]. \]

This means that the expectation of the gradient energy of a signal increases after data embedding. We can also conclude from Equations (1) and (8) that a more distinguishable \( GE^\alpha \) can be obtained if the amplitudes of \( r^\alpha_w(n) \) are enlarged by a larger \( \alpha \) or more secret date are hidden.

### 3.2 Coefficient distribution modification

The embedding scheme can be expressed as

\[ r^\alpha = r + J \cdot \alpha \cdot w \quad \text{or} \quad r^\bar{A} = r'(1 + J \cdot \alpha \cdot w) \] (9)

where \( J \) is the corresponding masking threshold (if any) based on the visual model.

Let \( m \) be the amount of disturbance \( (m = J \cdot \alpha \cdot w \) or \( m = r' \cdot J \cdot \alpha \cdot w \), based on Eq. (9)) added to coefficients. Since the embedded data \{ \( w \) \} is generally a noise-like sequence of random polarity after scrambling, it is reasonable to assume that \( m \) is also noise-like sequence with zero mean, i.e., \( E[m] = 0 \), and \( m \) is statistically independent of \( r^\alpha \). Then, we can compute the statistical properties for \( A \),

\[ \mu^\alpha = E[r^\alpha] = \mu^\alpha \quad \text{and} \quad \text{Var}[r^\alpha] = \text{Var}[r^\alpha] + \sigma^2 \]

where \( \text{Var}[m] = \sigma^2 \) is the variance of \( m \), \( \mu^\alpha \) and \( \mu^\alpha \) represent coefficient means of the original and the stego version, respectively.

In the same manner, because coefficients within \( \bar{A} \) are not used to convey secret data, we can obtain \( \mu^\bar{A} = \mu^\bar{A} \) and \( \text{Var}[r^\bar{A}] = \text{Var}[r^\bar{A}] \), where \( \mu^\bar{A} \) and \( \text{Var}[r^\bar{A}] \) are means and variances of the original coefficients and stego-one’s, respectively. Then, we can get the mean of stego coefficients, \( \mu^\bar{A} \), as

\[ \mu^\bar{A} = \frac{N_1}{N_1+N_2} \mu^\alpha + \frac{N_2}{N_1+N_2} \mu^\bar{A} + \mu^\bar{A}. \] (10)

where \( \mu^\bar{A} \) stands for the mean of original coefficients. The variance \( \text{Var}[r^\bar{A}] \) can be computed as

\[ \text{Var}[r^\bar{A}] = \frac{N_1}{N_1+N_2} \text{Var}[r^\alpha] + \frac{N_2}{N_1+N_2} \text{Var}[r^\bar{A}] \]

\[ + \frac{N_1(\mu^\bar{A} - \mu^\alpha)^2 + N_2(\mu^\bar{A} - \mu^\alpha)^2}{N_1+N_2}. \] (11)

We also compute the variance \( \text{Var}[r^\bar{A}] \) as

\[ \text{Var}[r^\bar{A}] = \frac{N_1}{N_1+N_2} \text{Var}[r^\alpha] + \frac{N_2}{N_1+N_2} \text{Var}[r^\bar{A}] \]

\[ + \frac{N_1(\mu^\bar{A} - \mu^\alpha)^2 + N_2(\mu^\bar{A} - \mu^\alpha)^2}{N_1+N_2}. \] (12)

Based on Eqs. (11) and (12), we obtain

\[ \text{Var}[r^\bar{A}] - \text{Var}[r^\alpha] = (N_i / (N_i + N_2)) \sigma^2. \]

Since \( \sigma^2 \geq 0 \), we have: \( \text{Var}[r^\bar{A}] \geq \text{Var}[r^\alpha] \). It means that the alteration of the variance of coefficients will be evident as secret data with larger energy are embedded.

### 4. Experimental Results

Several public-domain embedding methods based on two techniques (i.e. LSB embedding and spread spectrum) were chosen for testing. They can be categorized into data embedding techniques in the pixel, the DCT, and the DWT domains. We selected 30 images of size 256×256 as the host image and utilized them to generate 90 non-stego and 90 stego images for evaluation in each domain.
The discriminant power of selected features is first analyzed. A set of normalized feature vectors, \( \mathbf{f}_i^k = \mathbf{f}_i^k / \sigma_i^k \), \( i=1,2, \ k=\text{Pixel, DCT, DWT} \) are calculated for each seed image to study relative feature variations after and before data hiding. In Table 1, we show the percentage of images which have a value deviates from 1 with a significant amount, in which 180 images including innocent and stego versions are used to evaluate the performance of the classifier in each domain. As shown in Table 1, all features have been modified significantly so that they are all suitable for steganalysis. Furthermore, we show the distribution of original and stego images in the 2-D feature space with respect to three embedding domains. The stego class and the original class are indicated by “o” and “●”, respectively. It is clear that the two classes are well separated except for those in the DWT domain. This implies that selected features are useful to split the stego class from the original one.

**Table 1.** Ratios for pixel, DCT, and DWT domains.

<table>
<thead>
<tr>
<th></th>
<th>( k=\text{P} )</th>
<th>( k=\text{D} )</th>
<th>( k=\text{W} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbf{f}_1^k )</td>
<td>97.78%</td>
<td>97.78%</td>
<td>85.56%</td>
</tr>
<tr>
<td>( \mathbf{f}_2^k )</td>
<td>97.78%</td>
<td>94.44%</td>
<td>98.89%</td>
</tr>
</tbody>
</table>

The classification rates by using the Baysian and multi-layer neural classifiers are given in Tables 2 and 3, and where PD, ND, FP, and FN represent Positive Detection, Negative Detection, False Positive, and False Negative, respectively. As shown in these two tables, the classification rates \( i.e. 0.5*(PD+ND) \) are higher than 82.2% except for the case where the Bayesian classifier is applied to the DWT domain features (79.4%). The multi-layer neural classifier outperforms the Bayesian classifier, since it supplies a non-linear decision boundary.

**Table 2.** Classification results (Bayesian classifier).

<table>
<thead>
<tr>
<th></th>
<th>PD</th>
<th>ND</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel domain</td>
<td>84.44%</td>
<td>83.33%</td>
<td>16.67%</td>
<td>15.56%</td>
</tr>
<tr>
<td>DCT domain</td>
<td>92.22%</td>
<td>77.78%</td>
<td>22.22%</td>
<td>7.78%</td>
</tr>
<tr>
<td>DWT domain</td>
<td>82.22%</td>
<td>76.67%</td>
<td>23.33%</td>
<td>17.78%</td>
</tr>
</tbody>
</table>

**Table 3.** Classification results (the neural classifier).

<table>
<thead>
<tr>
<th></th>
<th>PD</th>
<th>ND</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel domain</td>
<td>84.44%</td>
<td>95.56%</td>
<td>4.44%</td>
<td>15.56%</td>
</tr>
<tr>
<td>DCT domain</td>
<td>88.89%</td>
<td>94.44%</td>
<td>5.56%</td>
<td>11.11%</td>
</tr>
<tr>
<td>DWT domain</td>
<td>81.11%</td>
<td>83.33%</td>
<td>16.67%</td>
<td>18.89%</td>
</tr>
</tbody>
</table>

**Table 4.** Classification results between pixel and DCT domains with a neural classifier.

<table>
<thead>
<tr>
<th></th>
<th>Pixel Domain</th>
<th>DCT Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection Rate</td>
<td>86.67%</td>
<td>90%</td>
</tr>
<tr>
<td>Missing Rate</td>
<td>13.33%</td>
<td>10%</td>
</tr>
</tbody>
</table>

By concentrating on the discrimination of the pixel and the DCT domain data hiding, the classification rate using a multi-layer neural classifier is shown in Table 4. The result demonstrates that our scheme can achieve good hiding domain discrimination between DCT and pixel domains.

Fig. 1 The distribution of feature vectors in the (a) pixel, (b) DCT, and (c) DWT domains, where “●” and “o” represent the original image samples and stego ones, respectively.

5. References